Mixture Models for Image Analysis

Florence Forbes

16

Mixture Models for Image Analysis

Florence Forbes

Laboratoire Jean Kuntzmann, Université Grenoble Alpes and INRIA Grenoble, France

CONTENTS

16.1 Introduction

Image analysis includes a variety of tasks such as image restoration, segmentation, registration, visual tracking, retrieval, texture modelling, classification and sensor fusion. Important application domains are medical imaging, remote sensing and computer vision. Problems involving incomplete data, where part of the data is missing or unobservable, are common, and mixture models can be used in many of these tasks directly or indirectly. The aim may be to recover an original image which is hidden and has to be estimated from a noisy or blurred version (restoration). More generally, the observed and hidden data are not necessarily of the same nature (segmentation). The observations may represent measurements, for example, multidimensional variables recorded at each pixel of an image, while the hidden data could consist of an unknown class assignment to be estimated at each pixel. To give an idea of the variety of uses, mixture models have been used for image restoration (e.g. Niknejad et al., 2015), for image registration (Gerogiannis et al., 2009), visual tracking (Karavasilis et al., 2012), image retrieval (Beecks et al., 2015), texture modelling (Blanchet & Forbes, 2008), classification (Bouveyron et al., 2007) and sensor fusion (Gebru et al., 2016), to name only a few of the relevant papers. However, the most typical and direct use

relates to image segmentation which can be recast straightforwardly into a clustering task. More generally, in this chapter we will focus on problems that can be posed as labelling or clustering problems in which the solution is a set of labels assigned to image pixels or features.

In the context of statistical image segmentation, choosing the probabilistic model that best accounts for the observations is an important first step for the quality of the subsequent estimation and analysis. Hidden Markov random field (HMRF) models were revealed to be a powerful tool for image segmentation (Geman & Geman, 1984; Besag, 1986). They are very useful in accounting for spatial dependencies between the different pixels of an image, but these spatial dependencies are also responsible for a typically large amount of computation. Markov model-based segmentation requires estimation of the model parameters. A common approach involves alternately restoring the unknown segmentation (labelling or clustering) based on a maximum *a posteriori* rule and then estimating the model parameters using the observations and the restored data. This is the case, for instance, in the popular iterated conditional mode (ICM) algorithm of Besag (1986) which makes use of the pseudo-likelihood approximation (Besag, 1974). This combination usually provides reasonable segmentations but is known to lead to biased parameter estimates, essentially due to the restoration step. Because of the missing data structure of the task, the expectation-maximization (EM) algorithm provides another justifiable formalism for such an alternating scheme. It has the advantage of dealing with conditional probabilities instead of committing to suboptimal restorations of the hidden data.

In this chapter, we first present how HMRF models generalize standard mixture models (Section 16.2). We propose an inference procedure using variational approximation (Section 16.3) and illustrate the framework with two real medical image applications (Section 16.4).

16.2 Hidden Markov Model Based Clustering

Hidden structure models, and more specifically Gaussian mixture models, are among the most statistically mature methods for clustering. A clustering or labelling problem is specified in terms of a set of sites S and a set of labels $\mathcal{G}$. A site often represents an item, a point or a region in Euclidean space such as an image pixel or an image feature. A set of sites may be categorized in terms of their regularity. Sites on a lattice are considered as spatially regular (e.g. the pixels of a two-dimensional image). Sites which do not present spatial regularity are considered as irregular. This is the usual case when sites represent geographic locations (Green & Richardson, 2002) or features extracted from images at a more abstract level, such as *interest points* (see Lowe, 2004; Blanchet & Forbes, 2008). It can also be that the sites correspond to items (e.g. genes) that are related to each other through a distance or dissimilarity measure (Vignes & Forbes, 2009) or simply to a collection of independent items.

A label is an event that may happen to a site. We will consider only the case where a label assumes a discrete value in a set of G labels. In the following developments, it is convenient to consider $\mathcal{G}$ as the set of G-dimensional indicator vectors $\mathcal{G} = \{e_1, \ldots, e_G\}$, where each e_g has all its components being 0, except the gth which is 1. The labelling problem is to assign a label from a label set $\mathcal{G}$ to each of the sites. If there are n sites, the set $\mathbf{z} = \{z_1, \ldots, z_n\}$, with $z_i \in \mathcal{G}$ for all $i \in S$, is called a labelling of the sites in S in terms of the labels in $\mathcal{G}$. We consider cases where the data naturally divide into observed data $\mathbf{y} = \{y_1, \ldots, y_n\}$ and unobserved or missing membership data $\mathbf{z} = \{z_1, \ldots, z_n\}$. They are

considered as random variables denoted by $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ and $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ with domain $\mathcal{Z}$ being equal to $\mathcal{Z} = \mathcal{G}^n$.

For image analysis or spatial data clustering, dependencies or contextual information can be taken into account using HMRF models, which can be seen as a generalization of standard mixture models.

16.2.1 Mixture models

Before introducing the HMRF model, we recall the underlying mixture modelling corresponding to independent Z_i. The model reduces to a standard mixture model that we will refer to as an *independent mixture*; see also Chapter 1. The distribution of $(\mathbf{Y}, \mathbf{Z})$ is defined by

$$p(\mathbf{z}) = \prod_{i \in S} p(z_i), \tag{16.1}$$

$$p(\mathbf{y}|\mathbf{z}) = \prod_{i \in S} p(y_i|z_i). \tag{16.2}$$

Equation (16.1) means that the hidden variables Z_i are independent, while equation (16.2) is sometimes referred to as the *independent noise assumption*. Under (16.1) and (16.2), the Y_i are also independent variables. To recover the standard mixture definition, we need to assume that the Z_i are identically distributed according to a multinomial distribution with parameters $\eta = \{\eta_1, \ldots, \eta_G\}$. Similarly, the conditional distribution for class g, $p(\cdot|Z_i = e_g) = f(\cdot|\theta_g)$, is assumed not to depend on i but only on some parameter θ_g. Different choices are possible for $f(\cdot|\theta_g)$. The most commonly encountered in applications are multivariate Gaussians (Celeux & Govaert, 1995; Banfield & Raftery, 1993), multivariate Student (McLachlan & Peel, 2000; Gerogiannis et al., 2009), and Poisson distributions for count data (Green & Richardson, 2002; Forbes et al., 2013; Karlis & Meligkotsidou, 2007); see also Chapter 8 above for a review of mixture modelling of count data.

16.2.2 Markov random fields: Potts model and extensions

When the Z_i are not independent, the interrelationship between sites can be modelled by a so-called neighbourhood system usually defined through a graph. Two neighbouring sites correspond to two nodes of the graph linked by an edge. The dependencies between neighbouring Z_i are then modelled by further assuming that the joint distribution of $Z_1, \ldots, Z_n$ is a discrete Markov random field (MRF) on this specific graph defined by

$$p(\mathbf{z}) = W^{-1} \exp(-H(\mathbf{z})), \tag{16.3}$$

where W is a normalizing constant and H is a function assumed to be of the following form (we restrict to pairwise interactions):

$$H(\mathbf{z}) = \sum_{i \sim j} V_{ij}(z_i, z_j) + \sum_{i \in S} V_i(z_i), \tag{16.4}$$

where the V_{ij} (V_i) are functions referred to as pair (singleton) potentials. We write $i \sim j$ when sites i and j are neighbours on the graph, so that the sum above is only over neighbouring sites.

A simple model is the so-called Ising model where the Z_i are binary variables representing spin orientations. The more general Potts model allows the Z_i to take G values that correspond to G classes with $G > 2$.

The singleton potentials $V_i(z_i)$ impact the probability of assigning site i to label or class z_i. When these potentials depend on i and not only on z_i, they are referred to as a non-stationary external field. Such non-stationarity can be useful to account for *a priori* knowledge that may vary with the site. This is typically the case when introducing probabilistic atlases in brain MRI data analysis (Forbes et al., 2011). Often, however, we restrict to a stationary external field that can be then denoted by $V_i(z_i) = -\alpha_{z_i}$. The potentials are then defined by a G-dimensional vector $\alpha = \{\alpha_1, \ldots, \alpha_G\}$ using the vector notation $V_i(z_i) = -\langle z_i, \alpha \rangle$, where $\langle z_i, \alpha \rangle$ denotes the scalar product between z_i and α. The latter notation has the advantage that it still makes sense when the vectors are arbitrary and not necessarily indicators. This will be useful when describing the algorithms of Section 16.3.

The singleton potentials are linked to the standard mixture proportions η. When the V_{ij} are zero in (16.4), $p(\mathbf{z})$ in (16.3) reduces to a standard mixture model up to the reparameterization

$$p(\mathbf{z}) \quad = \quad \prod_{i \in S} \frac{\exp(\langle z_i, \alpha \rangle)}{\sum_{g'=1}^{G} \exp(\alpha_{g'})} \, . \tag{16.5}$$

From (16.5), we can identify the link between α and η as, for all $g = 1, \ldots, G$,

$$\eta_g \quad = \quad \frac{\exp(\alpha_g)}{\sum_{g'=1}^{G} \exp(\alpha_{g'})}.$$

The pair potentials allow us to model the dependence between Z_i and Z_j at sites i and j. We consider pair potentials V_{ij} that depend on z_i and z_j but also possibly on i and j. Since the z_i can only take a finite number of values, for each i and j, we can define a $G \times G$ matrix $\mathbb{V}_{ij} = (\mathbb{V}_{ij}(k,l))_{1 \leq k,l \leq G}$ and write without lost of generality $V_{ij}(z_i, z_j) = -\mathbb{V}_{ij}(k,l)$ if $z_i = e_k$ and $z_j = e_l$, or, using the indicator vector notation, $V_{ij}(z_i, z_j) = -\langle z_i, \mathbb{V}_{ij} z_j \rangle$.

If, for all i and j, $\mathbb{V}_{ij} = \beta \times I_G$ where β is a scalar and I_G is the $G \times G$ identity matrix, then the pair potentials reduce to a single scalar interaction parameter β and we get the Potts model traditionally used for image segmentation (Besag, 1986). Note that this model is appropriate most of the time for segmentation since, for positive β, it tends to favour neighbours that are in the same class.

In practice, these parameters can be tuned according to experts or *a priori* knowledge or they can be estimated from the data. In the latter case, the part to be estimated is usually assumed independent of the indices i and j. In what follows, the Markov model parameters will reduce to a single matrix $\mathbb{V}$. Note that, formulated as such, the model is not identifiable in the sense that different values of the parameters, namely $\mathbb{V}$ and $\mathbb{V} + c\mathbb{1}$ (where $\mathbb{1}$ denotes the $G \times G$ matrix with all its components being 1 and c an arbitrary scalar value) lead to the same probability distribution. This issue is generally easily handled by imposing some additional constraint such as $\mathbb{V}(k,l) = 0$ for one of the components (k,l).

16.2.3 Hidden Markov field with independent noise

The independent noise assumption (16.2) underlying standard mixture models is also crucial in the more general hidden Markov random field. When the goal is to estimate $\mathbf{z}$ from the observed $\mathbf{Y} = \mathbf{y}$, most approaches fall into two categories. The first ones focus on finding the best $\mathbf{z}$ using a Bayesian decision principle such as maximum *a posteriori* or maximum posterior mode rules. This explicitly involves the use of $p(\mathbf{z}|\mathbf{y})$ and uses the fact that the conditional field denoted by $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$ is a Markov field. This includes methods such as ICM (Besag, 1986) and simulated annealing (Geman & Geman, 1984) which differ in the way they deal with the intractable $p(\mathbf{z}|\mathbf{y})$ and use its Markovianity. A second type of approach is related to a missing-data point of view, for which the focus is on estimating parameters

when some of the data are missing (the z_i here). The reference algorithm in such cases is the EM algorithm (Dempster et al., 1977). In addition to estimating the parameters, the EM algorithm provides also a segmentation $\mathbf{z}$ by offering the possibility of restoring the missing data; see Chapter 2 above for a detailed review of the EM algorithm.

However, when applied to hidden Markov fields, the algorithm is not tractable and requires approximations. It follows a number of procedures including the Gibbsian EM of Chalmond (1989), the MCEM algorithm and a generalization of it (Qian & Titterington, 1991), the PPL-EM algorithm of Qian & Titterington (1991) and various mean-field-like approximations of EM (Celeux et al., 2003). Such approximations are also all based on the Markovianity of $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$. This property is a critical requirement for any further developments.

When $\mathbf{Z}$ is Markovian, a simple way to guarantee the Markovianity of $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$ is the independent noise assumption (16.2). Indeed, equations (16.2) and (16.3) imply that the conditional field $(\mathbf{Y}, \mathbf{Z})$ is a Markov random field, which implies that $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$ is an MRF too. This standard and widely used situation is referred to in Benboudjema & Pieczynski (2005) as the hidden Markov field with independent noise (HMF-IN) model. Equation (16.2) is a conditional independence and non-correlated noise condition. Denoting by $\theta = \{\theta_1, \ldots, \theta_G\}$ the class-dependent distribution parameters, the HMF-IN parameters are denoted by $\Psi = (\theta, \alpha, \mathbb{V})$. In the one-dimensional Gaussian case, $\theta_g = (\mu_g, \sigma_g^2)$, the mean and variance parameters of the Gaussian distribution.

Like standard mixture models, hidden Markov (random) fields can then be used for a number of segmentation or clustering tasks. Many applications are related to image analysis, but other examples include population genetics (François et al., 2006) and bioinformatics (Vignes & Forbes, 2009). The fact that $\mathbf{Z}$ is Markovian is not strictly necessary. However, in a segmentation or clustering context, it has the advantage of providing some insight into and control of the segmentation regularity through a meaningful and easy-to-understand parametric model, but it also somewhat reduces the modelling capabilities of the approach (see Blanchet & Forbes, 2008). More general approaches involve so-called couple MRF or triplet MRF (Benboudjema & Pieczynski, 2005; Blanchet & Forbes, 2008) but will not be described in this chapter.

16.3 Markov Model Based Segmentation via Variational EM

The model complexity of a hidden Markov random field is greater than that of standard mixtures and makes the EM algorithm intractable. Solutions have been proposed which associate the pseudo-likelihood approximation (Besag, 1974) and Monte Carlo simulations (Chalmond, 1989), but the corresponding algorithms are time-consuming. In this section we present a variational approximation approach. In a number of complex real imaging applications, it has been observed as a competitive alternative to Markov chain Monte Carlo approaches, in terms of the quality of the results, with a great gain in terms of computation time (for a comparison in functional MRI analysis, see, for example, Chaari et al., 2013). The variational approximation relates to the so-called mean field approximation in statistical physics (Chandler, 1987). For a hidden Markov field model, the likelihood of $(\mathbf{Y}, \mathbf{Z})$ is called the complete (or complete-data) likelihood and is given by

$$p(\mathbf{y}, \mathbf{z}|\Psi) = p(\mathbf{y}|\mathbf{z}, \theta)\, p(\mathbf{z}|\alpha, \mathbb{V}). \tag{16.6}$$

The conditional field $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ is also a Markov field, with energy function $H(\mathbf{z}; \alpha, \mathbb{V}) - \log p(\mathbf{y}|\mathbf{z}, \theta)$. Henceforth, we will refer to the Markov fields $\mathbf{Z}$ and $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ as the

marginal and the conditional fields, respectively. Recovering the unknown $\mathbf{Z}$ requires values for the vector parameter $\Psi = (\theta, \alpha, \mathbb{V})$. If unknown, the parameters are often estimated from the maximum likelihood perspective as

$$\hat{\Psi} = \arg\max_{\Psi} \log p(\mathbf{y}|\Psi), \tag{16.7}$$

where $p(\mathbf{y}|\Psi)$ is the incomplete (also called observed-data) likelihood. This optimization is usually solved by the iterative EM procedure (Dempster et al., 1977); see also Chapter 2 above. Each iteration may be formally decomposed into two steps. Given the current value of the parameter $\Psi^{(s)}$ at iteration s, the E step involves computing the expectation of the complete log likelihood knowing the observations $\mathbf{y}$ and the current estimate $\Psi^{(s)}$. In the M step, the parameter is then updated by maximizing this expected complete log likelihood,

$$\Psi^{(s+1)} = \arg\max_{\Psi} \sum_{\mathbf{z}\in\mathcal{Z}} \log p(\mathbf{y}, \mathbf{z}|\Psi)\ p(\mathbf{z}|\mathbf{y}, \Psi^{(s)}). \tag{16.8}$$

It is known that, under mild regularity conditions, EM converges to the set of stationary points of the incomplete likelihood $\Psi \mapsto p(\mathbf{y}|\Psi)$ (Wu, 1983). As discussed in Csiszar & Tusnady (1984) and Neal & Hinton (1998), denoting by $\mathcal{D}$ the set of distributions on missing data, EM can be viewed as an alternating maximization procedure of a function F defined, for any probability distribution $q \in \mathcal{D}$, by

$$F(q, \Psi) = \sum_{\mathbf{z}\in\mathcal{Z}} \log\left(\frac{p(\mathbf{y}, \mathbf{z}|\Psi)}{q(\mathbf{z})}\right)\ q(\mathbf{z}). \tag{16.9}$$

Starting from current values $(q^{(s)}, \Psi^{(s)})$, set

$$q^{(s+1)} = \arg\max_{q\in\mathcal{D}}\ F(q, \Psi^{(s)}) \tag{16.10}$$

and

$$\begin{aligned}\Psi^{(s+1)} &= \arg\max_{\Psi}\ F(q^{(s+1)}, \Psi) \\ &= \arg\max_{\Psi}\ \sum_{\mathbf{z}\in\mathcal{Z}} \log p(\mathbf{y}, \mathbf{z}|\Psi)\ q^{(s+1)}(\mathbf{z}).\end{aligned} \tag{16.11}$$

The first optimization (16.10) has an explicit solution $q^{(s+1)} = p(\cdot|\mathbf{y}, \Psi^{(s)})$, so that the solutions of (16.8) and (16.11) are the same. Hence the "marginal" sequence $\{\Psi^{(s)}\}_s$ of the sequence $\{(q^{(s)}, \Psi^{(s)})\}_s$ produced by the alternating maximization procedure of F is an EM path. The maximization (16.11) can also be understood as the minimization of a Kullback–Leibler divergence, up to some convention on $p(\mathbf{y})$, thus justifying the name of alternating minimization procedure (e.g. Csiszar & Tusnady, 1984; Byrne & Gunawardana, 2005).

There exist different generalizations of EM when the M step (16.8) is intractable; it can be relaxed by requiring just an increase rather than an optimum. This yields generalized EM (GEM) procedures (McLachlan & Krishnan, 2008; see also Boyles, 1983, for a convergence result). Unfortunately, EM (or GEM) is not appropriate for solving the optimization problem (16.7) in HMRFs due to the complex structure of the hidden variables $\mathbf{Z}$. The distribution $p(\mathbf{z}|\alpha, \mathbb{V})$ is only known up to its normalizing constant W (the partition function) which depends upon the parameters of interest $\mathbb{V}$. The domain $\mathcal{Z}$ is too large, so that the E step is intractable.

Alternative approaches have been proposed and they can be understood as generalizations of the alternating maximization procedures mentioned above: the optimization (16.10)

is solved over a restricted class of probability distributions $\tilde{\mathcal{D}}$ on $\mathcal{Z}$, and the M step (16.11) remains unchanged. This yields the variational EM (VEM) algorithms (Jordan et al., 1998). For a convex optimization justification see also Wainwright & Jordan (2003, 2005). Byrne & Gunawardana (2005) proved that, under mild regularity conditions, VEM converges to the set $\mathcal{L}$ of the stationary points of the function F in $\tilde{\mathcal{D}}$. Here again, generalizations of VEM can be defined by requiring an increase rather than an optimum in the M step (16.11), thus defining generalized VEM procedures. These relaxation methods are part of the generalized alternating minimization procedures (Byrne & Gunawardana, 2005). The most popular form of VEM occurs when $\tilde{\mathcal{D}}$ is the set of independent probability distributions on $\mathcal{Z}$ so that $q^{(s+1)}(\mathbf{z})$ is a factorized distribution $\prod_{i\in S} q_i^{(s+1)}(z_i)$. Then optimizing (16.10) with respect to $q_i^{(s+1)}(e_k)$ leads to a fixed point equation for all $i \in S$ and for all $e_k \in V$:

$$\log q_i^{(s+1)}(e_k) = c_i + \sum_{\mathbf{z}\in\mathcal{Z}} \log p(\mathbf{z}|\mathbf{y}, \Psi^{(s)}) \mathbb{I}(z_i = e_k) \prod_{j\neq i} q_j^{(s+1)}(z_j), \tag{16.12}$$

where c_i is the normalizing constant. The Markov property implies that the right-hand side of the equation only involves the probability distributions q_j for j in the neighbourhood of i that we will denote by $j \in \mathcal{N}(i)$. Another equivalent form of (16.12) is to update in turn, for each i in S,

$$q_i^{(s+1)}(z_i) \propto \exp(\mathrm{E}_{q_{\setminus i}^{(s+1)}}[\log p(z_i|\mathbf{y}, \mathbf{Z}_{\setminus i}, \Psi^{(s)})]) \tag{16.13}$$

where the expectation is taken with regard to

$$q_{\setminus i}^{(s+1)}(\mathbf{z}_{\setminus i}) = \prod_{j\in\mathcal{N}(i)} q_j^{(s+1)}(z_j).$$

See Chaari et al. (2013, Appendix) for a straightforward way to derive (16.13) using the Kullback–Leibler divergence properties.

In practice, when developing the right-hand side of (16.13), the terms that do not depend on z_i can be omitted. The latter are part of the normalizing constant that can be deduced (e.g. in the exponential family case as explained in Beal & Ghahramani, 2003) or computed afterwards. In the HMF-IN case, it becomes

$$q_i^{(s+1)}(z_i) \propto \exp(\mathrm{E}_{q_{\setminus i}^{(s+1)}}[\log p(y_i|z_i, \theta^{(s)}) + \log p(z_i|\mathbf{Z}_{\setminus i}, \alpha^{(s)}, \mathbb{V}^{(s)})]).$$

For a pairwise potential MRF, it becomes

$$q_i^{(s+1)}(z_i) \propto p(y_i|z_i, \theta^{(s)}) \exp\left(\left\langle z_i, \sum_{j\in\mathcal{N}(i)} \mathbb{V}_{ij}^{(s)} \mathrm{E}_{q_j^{(s+1)}}(Z_j) + \alpha_i^{(s)} \right\rangle\right). \tag{16.14}$$

By way of illustration, let us consider a two-class Potts model for which $z_i \in \{e_1, e_2\}$ and the potentials are defined by $\alpha = 0$ and $V_{ij}(z_i, z_j) = \beta\langle z_i, z_j\rangle$, that is, $\mathbb{V}_{ij} = \mathbb{V} = \beta I_2$ where I_2 is the 2×2 identity matrix. It follows that $\mathrm{E}_{q_j^{(s+1)}}(Z_j) = (q_j^{(s+1)}(e_1), q_j^{(s+1)}(e_2))^\top$. Then equation (16.14) for $z_i = e_1$ reads

$$q_i^{(s+1)}(e_1) \propto p(y_i|z_i = e_1, \theta_1^{(s)}) \exp\left(\beta \sum_{j\in\mathcal{N}(i)} q_j^{(s+1)}(e_1)\right),$$

which after normalization leads to

$$\begin{aligned} q_i^{(s+1)}(e_1) &= \left(1 + \frac{p(y_i|z_i = e_2, \theta_2^{(s)})}{p(y_i|z_i = e_1, \theta_1^{(s)})} \exp\left\{\beta \sum_{j \in \mathcal{N}(i)} (q_j^{(s+1)}(e_2) - q_j^{(s+1)}(e_1))\right\}\right)^{-1}, \\ q_i^{(s+1)}(e_2) &= 1 - q_i^{(s+1)}(e_1). \end{aligned} \tag{16.15}$$

The fixed point equations (16.15) must be solved iteratively, updating each site in turn.

Equation (16.14) can also be recovered from a different point of view. The idea when considering a particular site i is to neglect the fluctuations of the sites interacting with i so that the resulting system behaves as one composed of independent variables. More specifically, for all j different from i, the z_j are fixed at their current conditional mean value $\mathrm{E}(Z_j|\mathbf{y}, \Psi^{(s)}])$. However, these mean values are unknown and it is the goal of the approximation to compute them. Therefore, the approximation depends on a self-consistency condition: the mean values that can be computed from the approximate distribution must be equal to the mean values used to define this approximate distribution. Then replacing the exact conditional mean values by the mean values in the approximation leads to a fixed point equation involving these mean values (see Celeux et al., 2003, for more details). Existence and uniqueness of a solution to (16.12) are properties that have not yet been fully understood and will not be discussed here. We refer to Tanaka (2001) for a better insight into the properties of the (potentially multiple) solutions of the mean field equations. Such solutions are usually computed iteratively (see Ambroise & Govaert, 1998, Zhang, 1996, and an erratum in Fessler, 1998).

Despite the relaxation which may make the summation of the VEM E step explicit for a convenient choice of $\tilde{\mathcal{D}}$ (i.e. the computation of $F(q^{(s+1)}, \Psi)$ in (16.11)), VEM remains intractable for hidden Markov random fields. From (16.6) and (16.11), θ and $(\alpha, \mathbb{V})$ are updated independently, given $q^{(s+1)}$. Under additional commonly used assumptions on $p(\mathbf{y}|\mathbf{z}, \theta)$, $\theta^{(s+1)}$ is computed in closed form (see, for example, Section 16.4). The issue is the update of $(\alpha, \mathbb{V})$ since it requires an explicit expression of the partition function or some related quantities (its gradient, for example).

To overcome this difficulty, different approaches have been proposed. The *mean field* and *simulated field* algorithms proposed in Celeux et al. (2003) are alternatives to VEM that propagate the approximation $q^{(s+1)}$ of $p(\mathbf{z}|\mathbf{y}, \Psi^{(s)})$ to $p(\mathbf{z}|\alpha, \mathbb{V})$. The MCVEM approach (Forbes & Fort, 2007) differs from the previous one in that the approximation method does not lead to a simple valid model but appears as a succession of approximations to overcome successive computational difficulties. Similar ideas have been used successfully to estimate $\mathbb{V}$ in various applications, for example, in Chaari et al. (2013) and Forbes et al. (2013). Another common solution is to fixed $\mathbb{V}$ to a sequence of values using an annealing scheme; see, for example, Scherrer et al. (2009). The parameter α is often set to zero, although it can be added to the set of unknown parameters to be estimated without much difficulty (Celeux et al., 2004).

16.3.1 Links with the iterated conditional mode and the Gibbs sampler

As presented in Besag (1974), the iterated conditional mode algorithm involves updating in turn a solution $\mathbf{z}^*$ that satisfies

$$\begin{aligned} z_i^{*(s+1)} &= \arg\max_{z_i} p(y_i|z_i, \theta^{(s)}) p(z_i|\mathbf{z}_{\mathcal{N}(i)}^{*(s)}) \\ &= \arg\max_{z_i} p(y_i|z_i, \theta^{(s)}) \exp\left(\beta \left\langle z_i, \sum_{j \in \mathcal{N}(i)} z_i^{*(s)} \right\rangle\right). \end{aligned}$$

As noted in Celeux et al. (2003), it can be seen as a modal version of the variational mean field in the sense that the fixed point equations are similar, with the mean operator replaced by the mode (max) operator. Similarly, the Gibbs sampler as presented in Geman & Geman (1984) is recovered by sampling

$$z_i^{*(s+1)} \quad \sim \quad p(y_i|z_i,\theta^{(s)}) \exp\left(\beta \left\langle z_i, \sum_{j\in\mathcal{N}(i)} z_i^{*(s)} \right\rangle\right),$$

where $\sim$ indicates a simulation according to the distribution defined on the right-hand side. The Gibbs sampler can be seen as a simulated version of the mean field approximation.

16.4 Illustration: MRI Brain Scan Segmentation

We illustrate how the modelling and estimation scheme presented could provide general guidelines to deal with complex joint processes in medical image analysis. We provide two applications, both involving brain MRI data, but in different contexts and illustrating different capabilities of the models presented. Section 16.4.1 deals with image data where each pixel is associated with a univariate observation (a single MR sequence). The emphasis is on a sophisticated use of the external field or singleton potential parameters (α). In Section 16.4.2 multivariate observations are considered. Multiple MR sequences are segmented simultaneously. The emphasis is put on the design of the pair potential parameters ($\mathbb{V}$).

16.4.1 Healthy brain tissue and structure segmentation

The analysis of MR brain scans is a complex task that requires several sources of information to be taken into account and combined. The analysis is frequently based on segmentations of tissues and of subcortical structures performed by human experts. For automatic segmentation, difficulties arise from the presence of various artefacts such as noise or intensity non-uniformities (see Figure 16.1(a) and (c)). For structures, the segmentation requires in addition the use of prior information usually encoded via a pre-registered atlas. Interest has been growing (see, for example, Ashburner & Friston, 2005; Pohl et al., 2006) in tackling this complexity by allowing the possibility of introducing mutual interactions between components of a model. Such a coupling can be naturally expressed in a statistical framework via the definition of a joint distribution that performs a number of essential tasks. The statistical framework illustrated in this section allows (1) for tissue segmentation using *local* HMRF models, (2) for MRF segmentation of structures and (3) for *local affine* registration of an atlas. All tasks are linked, and completing each one of them can help in refining the others. We specify a joint model from which conditional models are derived. As a result, cooperation between tissues and structures and interaction between the segmentation and registration steps are easily introduced. An explicit joint formulation has the advantage of providing a strategy to construct more consistent or complete models that are open to incorporation of new tasks. Estimation is then carried out using a variational EM framework (see Scherrer et al., 2009, and Forbes et al., 2011, for details). The evaluation performed on both phantoms and real 3 tesla brain scans shows good results and demonstrates the clear improvement provided by coupling the registration step to tissue and structure segmentation.

16.4.1.1 A Markov random field approach to segmentation and registration

We consider a finite set S of n voxels on a regular three-dimensional grid. The observed data $\mathbf{y} = \{y_1, \ldots, y_n\}$ are the intensity values observed respectively at each voxel and the missing data $\mathbf{z} = (\mathbf{t}, \mathbf{s})$ is made up of two sets: the tissue classes $\mathbf{t} = \{t_1, \ldots, t_n\}$ and the subcortical structure classes $\mathbf{s} = \{s_1, \ldots, s_n\}$. The t_i take their values in $\{e_1, e_2, e_3\}$, which represents the three tissues cerebro-spinal fluid (CSF), grey matter and white matter (see Figure 16.1(b)). For the subcortical structure (see Figure 16.1(c)) segmentation we consider L structures, the s_i taking their values in $\{e'_1, \ldots, e'_L, e'_{L+1}\}$ where e'_{L+1} corresponds to an additional background class. Tissues and structures are linked and we denote by T^{s_i} the tissue of structure s_i at voxel i. The model parameters $\Psi = (\theta, \mathcal{R})$ include both the intensity distribution parameters θ and the registration parameters $\mathcal{R}$. We consider them in a Bayesian framework as realizations of random variables. The MRF parameters will be considered here as fixed (see below).

To capture interactions between the various fields $\mathbf{y}$, $\mathbf{t}$, $\mathbf{s}$ and Ψ we adopt a *conditional random field* approach which involves specifying a conditional model $p(\mathbf{t}, \mathbf{s}, \Psi|\mathbf{y})$. We define $p(\mathbf{t}, \mathbf{s}, \Psi|\mathbf{y})$ as a Gibbs measure with energy function $H(\mathbf{t}, \mathbf{s}, \Psi|\mathbf{y})$,

$$p(\mathbf{t}, \mathbf{s}, \Psi|\mathbf{y}) \propto \exp(H(\mathbf{t}, \mathbf{s}, \Psi|\mathbf{y})),$$

where the energy is decomposed into the following terms. We denote by $f(y_i|t_i, s_i, \theta_i)$ positive functions of y_i and consider the decomposition

$$\begin{aligned} H(\mathbf{t}, \mathbf{s}, \Psi|\mathbf{y}) &= H_T(\mathbf{t}) + H_S(\mathbf{s}) + H_{T,S}(\mathbf{t}, \mathbf{s}) + H_{T,\mathcal{R}}(\mathbf{t}, \mathcal{R}) + H_{S,\mathcal{R}}(\mathbf{s}, \mathcal{R}) \\ &\quad + H_\theta(\theta) + H_{\mathcal{R}}(\mathcal{R}) + \sum_{i \in S} \log f(y_i|t_i, s_i, \theta_i) \ . \end{aligned} \tag{16.16}$$

In what follows, we discuss a number of essential tasks and show the terms in (16.16) can be specified so that the model performs the tasks listed below.

Robust-to-noise segmentation

Robust-to-noise segmentation is generally addressed via MRF modelling. It introduces local spatial dependencies between voxels, providing a labelling regularization. For tissue and structure segmentations, we use the standard Potts model setting

$$\begin{aligned} H_T(\mathbf{t}) &= \sum_{i \in S} \sum_{j \in \mathcal{N}(i)} \beta_T \langle t_i, t_j \rangle, \\ H_S(\mathbf{s}) &= \sum_{i \in S} \sum_{j \in \mathcal{N}(i)} \beta_S \langle s_i, s_j \rangle, \end{aligned}$$

where $\langle t_i, t_j \rangle$ denotes the scalar product, $\mathcal{N}(i)$ represents the voxels neighbouring i, and β_T and β_S are additional interaction strength parameters.

Local approach to deal with non-uniformity

Tissue intensity models are generally estimated globally through the entire volume and then suffer from imperfections at a local level. We adopt a local segmentation alternative. The principle is to locally compute the tissue models in various subvolumes of the initial volume. These models better reflect local intensity distributions and are likely to handle different sources of intensity non-uniformity.

We consider intensity models that depend on the tissue class k but also on the voxel localization, so that θ decomposes into $\theta = \{\theta_i, \ i \in S\}$ where $\theta_i = (\theta_i^1, \theta_i^2, \theta_i^3)^\top$. Although

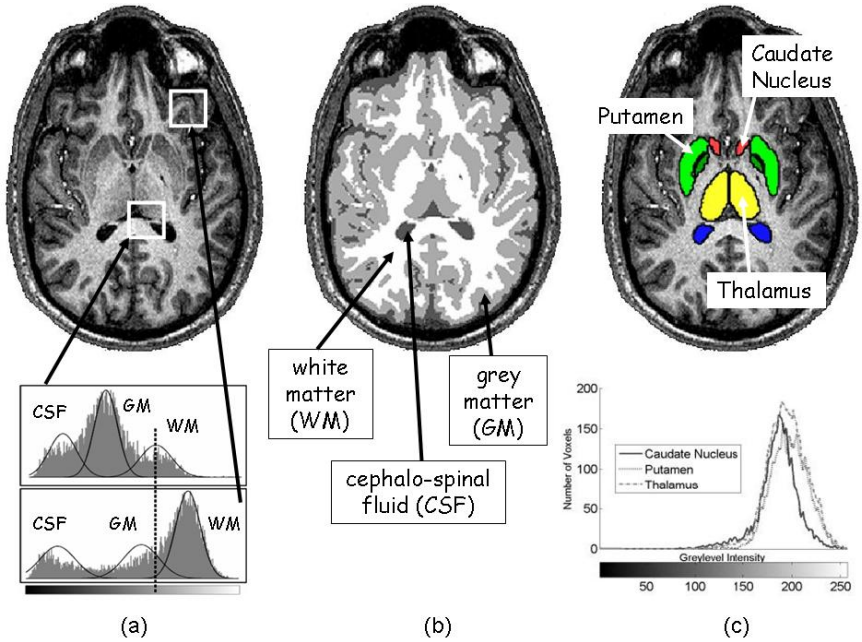


FIGURE 16.1
Obstacles to accurate segmentation of MR brain scans. Image (a) illustrates spatial intensity variations: two local intensity histograms (bottom) in two different subvolumes (top) are shown with their corresponding Gaussians fitted using three-component mixture models for the three brain tissues considered. The vertical line corresponds to some intensity value labelled as grey matter or white matter depending on the subvolume. Image (b) illustrates a segmentation into three tissues: white matter, grey matter and cerebro-spinal fluid. Image (c) shows the largely overlapping intensity histograms (bottom) of three grey matter structures segmented manually (top): the putamen, the thalamus and the caudate nuclei.

possible in our Bayesian framework, this general setting results in too many parameters which could not be estimated accurately. The local approach provides an intermediate efficient solution where the θ_i are first considered as constant over subvolumes. Let $\mathcal{C}$ be a regular cubic partitioning of the volume S into a number of non-overlapping subvolumes $\{V_c,\ c \in \mathcal{C}\}$. We write $\theta = \{\theta_c,\ c \in \mathcal{C}\}$, where $\theta_c = (\theta_c^1, \theta_c^2, \theta_c^3)^\top$ is the common value of all θ_i for $i \in V_c$. In addition, to ensure consistency and spatial regularity between the local estimations of the θ_cs, we consider an MRF prior $p(\theta) \propto \exp(H_\theta(\theta))$. When Gaussian intensity distributions are considered, this corresponds to assigning *auto-normal* Markov priors to the mean parameters. Apart from the issue of estimating θ, having voxel dependent θ_i is not a problem. We can easily return to this case from the estimated θ_cs, by using a cubic splines interpolation step.

Incorporating a priori *knowledge via local affine atlas registration*

The *a priori* knowledge required for structure segmentation is classically provided via a global non-rigid atlas registration. Most methods first register the prior information to the

medical image and then segment the image based on that aligned information. Although reliable registration methods are available, it is still important in the subsequent segmentation task to overcome biases caused by commitment to the initial registration. Also segmentation results provide information that can be used for feedback on registration. Global registration approaches generally lead to a high-dimensional minimization problem which is computationally greedy and subject to a large number of local optima.

We instead choose a hierarchical local affine registration model as in Pohl et al. (2006). We consider (a) a global affine transformation given by parameters $\mathcal{R}^G$, which describes the non-structure-dependent deformations, and (b) one local affine structure-dependent deformation for each structure, defined in relation to $\mathcal{R}^G$ and capturing the residual structure-specific deformations. It follows that $L+2$ affine transformation parameters $\mathcal{R} = (\mathcal{R}^G, \mathcal{R}^S_1, \ldots, \mathcal{R}^S_{L+1})$ have to be estimated. Interactions between labels and registration parameters are introduced through $H_{T,\mathcal{R}}(\mathbf{t}, \mathcal{R})$ and $H_{S,\mathcal{R}}(\mathbf{s}, \mathcal{R})$. Similarly to Pohl et al. (2006), the interaction between the structure classes $\mathbf{s}$ and $\mathcal{R}$ is chosen so as to favour configurations for which the segmentation of a structure l is aligned on its prior atlas. We denote by $\zeta_S = \{\zeta^l_S,\ l = 1, \ldots, L+1\}$ the statistical atlas of the brain subcortical structures under consideration and by $\mathfrak{f}(\mathcal{R}^G, \mathcal{R}^S_l, i)$ the interpolation function assigning a position in the atlas space to the image space. We compute the spatial *a priori* distribution $f^l_S(\mathcal{R}, \cdot)$ of one structure l as

$$f^l_S(\mathcal{R}, i) \quad = \quad \frac{\zeta^l_S(\mathfrak{f}(\mathcal{R}^G, \mathcal{R}^S_l, i))}{\sum_{l'=1}^{L+1} \zeta^{l'}_S(\mathfrak{f}(\mathcal{R}^G, \mathcal{R}^S_{l'}, i))} \ .$$

The normalization across all structures is necessary as $\mathcal{R}^S_l$ are structure-dependent parameters and multiple voxels in the atlas space could be mapped to one location in the image space. Although some atlas is potentially available for tissues, in our setting we build f_T, the spatial *a priori* distribution of the $K = 3$ tissues, from the f^l_S:

$$f^k_T(\mathcal{R}, i) \quad = \quad \sum_{l:T^l=k} f^l_S(\mathcal{R}, i) + \frac{1}{K} f^{L+1}_S(\mathcal{R}, i).$$

Agreement between structure segmentation and the atlas is then favoured by setting

$$H_{S,\mathcal{R}}(\mathbf{s}, \mathcal{R}) = \sum_{i \in S} \langle s_i, \log\left(f_S(\mathcal{R}, i) + \epsilon\right) \rangle,$$

with the vectorial notation $f_S = (f^1_S, \ldots, f^{L+1}_S)^\top$. The logarithm and a positive scalar ϵ are introduced respectively for homogeneity between probabilities and energies, and to ensure the existence of the logarithm. We choose $\epsilon = 1$, making in addition $H_{S,\mathcal{R}}(\mathbf{s}, \mathcal{R})$ positive, but the overall method does not seem to be sensitive to its exact value. Similarly, we define the interaction between $\mathbf{t}$ and $\mathcal{R}$ by

$$H_{T,\mathcal{R}}(\mathbf{t}, \mathcal{R}) = \sum_{i \in V} \langle t_i, \log\left(f_T(\mathcal{R}, i) + \epsilon\right) . \rangle$$

Then the term $H_{\mathcal{R}}(\mathcal{R})$ can be used to introduce *a priori* knowledge to favour estimation of $\mathcal{R}$ close to some average registration parameters computed from a training data set if available. In our case, no such data set were available and we set $H_{\mathcal{R}}(\mathcal{R}) = 0$.

Cooperative tissue and structure segmentations

Tissues and structures are linked: a structure is made up of a specific tissue and knowledge on structures, and locations provide information for tissue segmentation. Inducing cooperation

between tissue and structure segmentations can be done through the term $H_{T,S}(\mathbf{t}, \mathbf{s})$. We set

$$H_{T,S}(\mathbf{t}, \mathbf{s}) = \sum_{i \in S} \langle t_i, e_{T^{s_i}} \rangle$$

so as to favour situations for which the tissue T^{s_i} of structure s_i is the same as the tissue given by t_i. Cooperation between tissue and structure labels also appears via the energy data term $\sum_{i \in S} f(y_i | t_i, s_i, \theta_i)$. Considering Gaussian intensity distributions, we denote by $\mathcal{N}(\cdot | \mu, \lambda^{-1})$ the Gaussian distribution with mean μ and precision λ (i.e. the inverse of the variance). Denoting $\theta_i^k = \{\mu_i^k, \lambda_i^k\}$, we see θ_i as a three-dimensional vector, so that when $t_i = e_k$, $\mathcal{N}(y_i | \langle t_i, \theta_i \rangle)$ denotes the Gaussian distribution with mean μ_i^k and precision λ_i^k. To account for both tissue and structure information, we set

$$f(y_i | t_i, s_i, \theta_i) \quad = \quad \mathcal{N}(y_i | \langle t_i, \theta_i \rangle)^{\frac{1 + \langle s_i, e'_{L+1} \rangle}{2}} \mathcal{N}(y_i | \langle e_{T^{s_i}}, \theta_i \rangle)^{\frac{1 - \langle s_i, e'_{L+1} \rangle}{2}} .$$

When tissue and structure segmentations contain the same information at voxel i, that is, either $t_i = e_{T^{s_i}}$ or $s_i = e'_{L+1}$, the expression for f above reduces to the usual $\mathcal{N}(y_i | \langle t_i, \theta_i \rangle)$. When this is not the case, the expression for f above leads to $\mathcal{N}(y_i | \langle t_i, \theta_i \rangle)^{1/2} \mathcal{N}(y_i | \langle e_{T^{s_i}}, \theta_i \rangle)^{1/2}$, which is a more appropriate compromise.

This achieves the definition of the hierarchical model that can then be fitted to data using a VEM approach as specified in Scherrer et al. (2009) and Forbes et al. (2011).

16.4.1.2 Experiments: Joint tissue and structure segmentation

We consider both phantoms and real 3 T brain scans. We use the normal 1 mm^3 BrainWeb phantoms database from the McConnell Brain Imaging Center (Collins et al., 1998). These phantoms are generated from a realistic brain anatomical model and an MRI simulator that simulates MR acquisition physics, in which different values of non-uniformity and noise can be added. Because these images are simulated we can quantitatively compare our tissue segmentation to the underlying tissue generative model to evaluate the segmentation performance.

We perform a quantitative evaluation using the Dice similarity metric (Dice, 1945). This metric measures the overlap between a segmentation result and the gold standard. Denoting by TP_k the number of true positives for class k, FP_k the number of false positives and FN_k the number of false negatives, the Dice metric is given by

$$d_k \quad = \quad \frac{2\mathrm{TP}_k}{2\mathrm{TP}_k + \mathrm{FN}_k + \mathrm{FP}_k}.$$

It takes values in $[0, 1]$, where 1 represents perfect agreement. Since BrainWeb phantoms contain only tissue information, three subcortical structures were manually segmented by three experts: the left caudate nucleus, the left putamen and the left thalamus. The results we report are for eight BrainWeb phantoms, for 3%, 5%, 7% and 9% of noise with 20% and 40% of non-uniformity for each noise level. Regarding real data, we then evaluate our method on real 3 T MR brain scans (T1 weighted sequence) coming from the Grenoble Institute of Neuroscience.

We then evaluate the performance of the joint tissue and structure segmentation. We consider two cases: our combined approach with fixed registration parameters (LOCUSB-TS) and with estimated registration parameters (LOCUSB-TSR). Table 16.1 shows the evaluation on BrainWeb images using our reference segmentation of the three structures. The table shows the means and standard deviations of the Dice coefficient values obtained for the eight BrainWeb images. It also shows the means and standard deviations of the relative

TABLE 16.1
Mean Dice coefficient values obtained on three structures using LOCUSB-TS and LOCUSB-TSR for BrainWeb images, over eight experiments for different values of noise (3%, 5%, 7%, 9%) and non-uniformity (20%, 40%). The corresponding standard deviations are shown in parentheses. The second column shows the results when registration is done as a preprocessing step (LOCUSB-TS). The third columns shows the results with our full model including iterative estimation of the registration parameters (LOCUSB-TSR). The last column shows the relative Dice coefficient improvement for each structure.

Structure	LOCUSB-TS	LOCUSB-TSR	Relative Improvement
Left thalamus	91% (0)	94% (1)	4% (1)
Left putamen	90% (1)	95% (0)	6% (1)
Left caudate nucleus	74% (0)	91% (1)	23% (1)

improvements between the two models LOCUSB-TS and LOCUSB-TSR. In particular, a significant improvement of 23% is observed for the caudate nucleus for the latter model.

The three structure segmentations improve when registration is combined. In particular, in LOCUSB-TS the initial global registration of the caudate nucleus is largely suboptimal, but it is then corrected in LOCUSB-TSR. More generally, for the three structures we observe a stable gain for all noise and inhomogeneity levels.

Figure 16.2 shows the results obtained with LOCUSB-T, and LOCUSB-TSR on a real 3 T brain scan. The structures emphasized in image (c) are the two lateral ventricles, the caudate nuclei, the putamens and the thalamus. Figure 16.2(e) shows in addition a 3D reconstruction of 17 structures segmented with LOCUSB-TSR. The results with LOCUSB-TS are not shown because the differences with LOCUSB-TSR were not visible at this graphical resolution.

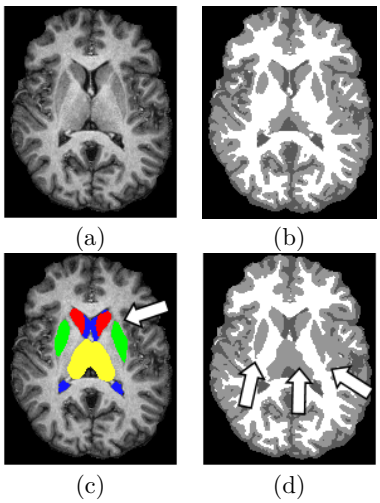
We therefore observe a gain in combining tissue and structure segmentation, in particular through the improvement of tissue segmentation for areas corresponding to structures such as the putamens and thalamus. The additional integration of a registration parameter estimation step also provides some significant improvement. It allows for an adaptive correction of the initial global registration parameters and a better registration of the atlas locally.

16.4.2 Brain tumor detection from multiple MR sequences

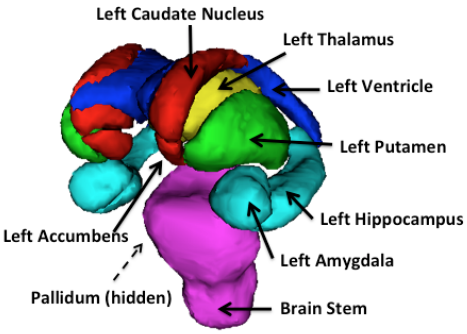
The previous subsection described a possible model for healthy brain segmentation using three normal tissues. When considering brain damage, the number of extra tissues to take into account can vary with the pathology. In this section, we illustrate the possibility of modelling interactions between these tissues via the pair potential parameters.

16.4.2.1 Tissue interaction modelling

A fully automatic algorithm is now proposed to segment glioma MR sequences, by availing of the additional information provided by multiple MR sequences. We adopt a data model comprising five normal tissue classes; white matter, grey matter, ventricular CSF, extraventricular CSF, and other. The glioma is modelled by a further four classes representing the diseased tissue state: oedema, non-enhancing, enhancing and necrotic. As illustrated in the previous section, the standard Potts model is often appropriate for clustering since it tends to favour neighbours that are in the same class. However, this model penalizes pairs that have different classes with the same penalty, regardless of the tissues they represent. In practice, it may be more appropriate to encode higher penalties when the tissues are known

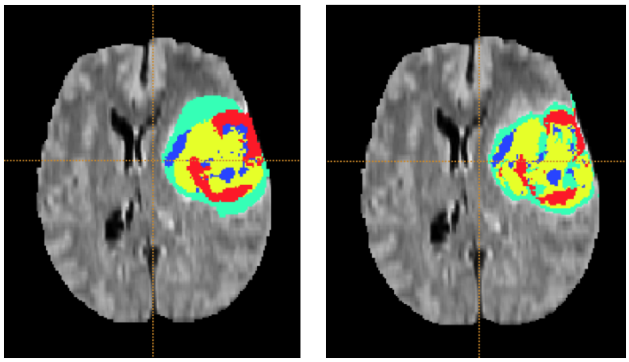


(a) (b) (c) (d)



(e)

FIGURE 16.2
Evaluation of LOCUSB-TSR on a real 3 T brain scan shown in image (a). For comparison, the tissue segmentation obtained with LOCUSB-TS is given in image (b). The results obtained with LOCUSB-TSR are shown in the second row. Major differences between tissue segmentations (images (c) and (d)) are indicated using arrows. Image (e) shows the corresponding 3D reconstruction of 17 structures segmented using LOCUSB-TSR. The names of the left structures (use symmetry for the right structures) are indicated in the image.



(a) (b)

FIGURE 16.3
Evaluation of P-LOCUS in image (b) on a real 3 T brain scan. The ground truth is shown in image (a).

to be unlikely neighbours. For example, the penalty for a white matter and extraventricular CSF pair is expected to be greater than that of a grey matter and extraventricular CSF pair, as these two classes are more likely to form neighbourhoods. This models the undesirability of abrupt changes in neighbouring tissues. In practice, the interaction matrix $\mathbb{V}$ can be tuned according to experts' *a priori* knowledge, or can be estimated from the data. In

the absence of sufficient data to robustly and accurately estimate a full free $\mathbb{V}$ with $G = 9$, further constraints are imposed on the matrix. The four glioma classes are considered a single structure, whose interaction with the normal tissue classes is not dependent on the specific glioma tissue state. Letting τ be the set of classes comprising the glioma structure and $\bar{\tau}$ the set of healthy tissues, we propose to use for $\mathbb{V}$ a matrix defined by

$$\mathbb{V}(g, g') = \begin{cases} \beta, & \text{for all } (g, g') \text{ such that } g \in \tau \text{ and } g' \in \bar{\tau}, \\ \beta_{g,g'}, & \text{otherwise.} \end{cases}$$

More generally, when prior knowledge indicates that, for example, two given classes are likely to be next to each other, this can be encoded in the matrix with a higher entry for this pair. Conversely, when there is enough information in the data, a full free $\mathbb{V}$ matrix can be estimated and will reflect the class structure (i.e. which class is next to which as indicated by the data) and will then mainly serve as a regularizing term to encode additional spatial information. The fine design of $\mathbb{V}$ may be important in such a case. For another illustration of a non-standard $\mathbb{V}$, see also Forbes et al. (2013).

For the distribution of the observed variables $\mathbf{y}$ given the classification $\mathbf{z}$, the usual conditional independence assumption is made. It follows that the conditional probability of the hidden field $\mathbf{z}$ given the observed field $\mathbf{y}$ is

$$p(\mathbf{z}|\mathbf{y}, \theta, \mathbb{V}) = W(\mathbb{V})^{-1} \exp\left(-H_{\mathbf{z}}(\mathbf{z}, \mathbb{V}) + \sum_{i \in S} \log f(y_i|z_i, \theta)\right).$$

For simplicity, no external field α is specified here, but it can be used in practice to account for prior knowledge via anatomical or vascular atlases (see Kabir et al., 2007, or the Appendix in Menze et al., 2015 for more details).

16.4.2.2 Experiments: Lesion segmentation

The algorithm referred to as P-LOCUS, for "Pathological LOCUS", was tested on real-patient data from the BRATS 2013 data set. A more complete description of the model used and the results is given in (Menze et al., 2015, Appendix). As an illustration, Figure 16.3 shows the correspondence between the ground truth corresponding to a manual segmentation and the P-LOCUS result.

16.5 Concluding Remarks

In this chapter we focused on image segmentation as a typical image processing task that can benefit from a mixture modelling approach. Regarding the specific brain MR application we described, the framework can be adapted to other applications. It provides a strategy and guidelines for dealing with complex joint processes involving more than one identified subprocess. It is based on the idea that defining conditional models is usually more straightforward and captures more explicitly cooperative aspects, including cooperation with external knowledge.

The Bayesian formulation provides additional flexibility such as the possibility of dealing, in a well-based manner, with some sort of non-stationarity in the parameters (like that due to intensity non-uniformities in our MRI example). Of course, depending on the application in mind, more complex energy functions than the one given in our MRI illustration may be necessary. In particular, for our example, it was enough to consider separately

cooperation between label sets and spatial interactions. However, one useful extension, to be investigated in future work, would be to add a spatial component in the cooperation mechanisms themselves.

Beyond this illustration, images can be considered in a broad sense meaning that the observed data do not need to be made up of a set of 2D or 3D pixels but could correspond to more general graph structures. The material in this chapter can be more generally applied to *dependent data clustering* as illustrated in Green & Richardson (2002), Vignes & Forbes (2009), and Forbes et al. (2013). In addition, we have not discussed a number of common complications that can occur in the measurement process. This includes issues such as the high dimensionality of the observations, missing observations and heterogeneous observations. Solutions exist in such cases. Some procedures using the EM approach are implemented in the SpacEM^3 software (Vignes et al., 2011) available at http://spacem3.gforge.inria.fr. Finally, since image analysis is a vast domain in terms of both methodology and applications, many important contributions are not cited or mentioned in this chapter.

Bibliography

Ambroise, C. & Govaert, G. (1998). Convergence proof of an EM-type algorithm for spatial clustering. *Pattern Recognition Letters* **19**, 919–927.

Ashburner, J. & Friston, K. J. (2005). Unified segmentation. *NeuroImage* **26**, 839–851.

Banfield, J. D. & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Beal, M. J. & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In *Bayesian Statistics 7*, J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds. Oxford: Oxford University Press.

Beecks, C., Uysal, M. S. & Seidl, T. (2015). Content-based image retrieval with Gaussian mixture models. In *MultiMedia Modeling: 21st International Conference, MMM 2015, Sydney,NSW, Australia, January 5-7, 2015, Proceedings, Part I*. Cham: Springer, pp. 294–305.

Benboudjema, D. & Pieczynski, W. (2005). Unsupervised image segmentation using triplet Markov fields. *Computer Vision and Image Understanding* **99**, 476–498.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B* **48**, 259–302.

Blanchet, J. & Forbes, F. (2008). Triplet Markov fields for the classification of complex structure data. *IEEE Transactions on Pattern Analysis Machine Intelligence* **30**, 1055–1067.

Bouveyron, C., Girard, S. & Schmid, C. (2007). High-dimensional discriminant analysis. *Communications in Statistics Theory and Methods* **36**, 2607–2623.

Boyles, R. (1983). On the convergence of EM algorithms. *Journal of the Royal Statistical Society, Series B* **45**, 47–50.

Byrne, W. & Gunawardana, A. (2005). Convergence theorems of generalized alternating minimization procedures. *Journal of Machine Learning Research* **6**, 2049–2073.

Celeux, G., Forbes, F. & Peyrard, N. (2003). EM procedures using mean field-like approximations for model-based image segmentation. *Pattern Recognition* **36**, 131–144.

Celeux, G., Forbes, F. & Peyrard, N. (2004). Modèle de Potts avec champ externe et algorithme de type EM pour la segmentation d'images. In *RFIA, 14th French Meeting AFRIF-AFIA Reconnaissance des Formes & Intelligence Artificielle*. Toulouse.

Celeux, G. & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* **28**, 781–793.

Chaari, L., Vincent, T., Forbes, F., Dojat, M. & Ciuciu, P. (2013). Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach. *IEEE Transactions on Medical Imaging* **32**, 821–837.

Chalmond, B. (1989). An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition* **22**, 747–761.

Chandler, D. (1987). *Introduction to Modern Statistical Mechanics*. New York: Oxford University Press.

Collins, D. L., Zijdenbos, A. P., Kollokian, V., Sled, J. G., Kabani, N. J., Holmes, C. J. & Evans, A. C. (1998). Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging* **17**, 463–468.

Csiszar, I. & Tusnady, G. (1984). Information geometry and alternating minimization procedures. *Statistics & Decisions* **1**, 205–237.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302.

Fessler, J. (1998). Comments on "The convergence of mean field procedures for MRF's". *IEEE Transactions on Image Processing* **7**, 917.

Forbes, F., Charras-Garrido, M., Azizi, L., Doyle, S. & Abrial, D. (2013). Spatial risk mapping for rare disease with hidden Markov fields and variational EM. *Annals of Applied Statistics* **7**, 1192–1216.

Forbes, F. & Fort, G. (2007). Combining Monte Carlo and mean field like methods for inference in hidden Markov random fields. *IEEE Transactions on Image Processing* **16**, 824–837.

Forbes, F., Scherrer, B. & Dojat, M. (2011). Bayesian Markov model for cooperative clustering: Application to robust MRI brain scan segmentation. *Journal de la Société Française de Statistique* **152**.

François, O., Ancelet, S. & Guillot, G. (2006). Bayesian clustering using hidden Markov random fields in spatial genetics. *Genetics* **174**, 805–816.

Gebru, I., Alameda-Pineda, X., Forbes, F. & Horaud, R. (2016). EM algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**, 2402–2415.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Gerogiannis, D., Nikou, C. & Likas, A. (2009). The mixtures of Student's t-distributions as a robust framework for rigid registration. *Image Vision Computing* **27**, 1285–1294.

Green, P. J. & Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of American Statistical Association* **97**, 1055–1070.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1998). An introduction to variational methods for graphical models. In *Learning in Graphical Models*, M. I. Jordan, ed. Dordrecht: Kluwer Academic Publishers, pp. 105–162.

Kabir, Y., Dojat, M., Scherrer, B., Forbes, F. & Garbay, C. (2007). Multimodal MRI segmentation of ischemic stroke lesions. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Lyon, France.

Karavasilis, V., Blekas, K. & Nikou, C. (2012). A novel framework for motion segmentation and tracking by clustering incomplete trajectories. *Computer Vision and Image Understanding* **116**, 1135–1148.

Karlis, D. & Meligkotsidou, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *Journal of Statistical Planning and Inference* **137**, 1942–1960.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **60**, 91–110.

McLachlan, G. & Krishnan, T. (2008). *The EM Algorithm and Extensions*. New York: Wiley. Second Edition.

McLachlan, G. & Peel, D. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* **10**, 339–348.

Menze, B., Reyes, M. & Van Leemput, K. e. a. (2015). The multimodal brain tumorimage segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34**, 1993–2024.

Neal, R. M. & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, M. I. Jordan, ed. Dordrecht: Kluwer Academic Publishers, pp. 355–368.

Niknejad, M., Rabbani, H. & Babaie-Zadeh, M. (2015). Image restoration using Gaussian mixture models with spatially constrained patch clustering. *IEEE Transactions on Image Processing* **24**, 3624–3636.

Pohl, K. M., Fisher, J., Grimson, E., Kikinis, R. & Wells, W. M. (2006). A Bayesian model for joint segmentation and registration. *NeuroImage* **31**, 228–239.

Qian, W. & Titterington, M. (1991). Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society London Series A* **337**, 407–428.

Scherrer, B., Forbes, F. & Dojat, M. (2009). A conditional random field approach for coupling local registration with robust tissue and structure segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009: 12th International Conference, London*, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble & C. Taylor, eds. Berlin: Springer-Verlag.

Tanaka, T. (2001). Information geometry of mean-field approximation. In *Advanced Mean Field Methods*, M. Opper & D. Saad, eds., chap. 17. Cambridge, MA: MIT Press.

Vignes, M., Blanchet, J., Leroux, D. & Forbes, F. (2011). SpaCEM3, a software for biological module detection when data is incomplete, high dimensional and dependent. *Bioinformatics* **27**, 881–882.

Vignes, M. & Forbes, F. (2009). Gene clustering via integrated Markov models combining individual and pairwise features. *IEEE Transaction on Computational Biology and Bioinformatics* **6**, 260–270.

Wainwright, M. & Jordan, M. (2003). Graphical models, exponential families, and variational inference. Tech. Rep. 649, UC Berkeley, Department of Statistics.

Wainwright, M. & Jordan, M. (2005). A variational principle for graphical models. In *New Directions in Statistical Signal Processing*, S. Haykin, T. Principe, T. Sejnowski & J. McWhirter, eds., chap. 11. Cambridge, MA: MIT Press.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.

Zhang, J. (1996). The convergence of mean field procedures for MRF's. *IEEE Transactions on Image Processing* **5**, 1662–1665.